The effect of speaker sampling to system stability in Likelihood Ratio based forensic voice comparison

xxxxx. xxxxxxx. Xxxx.

xxxx
{xxx/xxx/xxx}@xx.xx.xx

ABSTRACT

In forensic voice comparison (FVC) cases it is essential to make sure that conclusions are reliable, robust, and replicable. This is especially true for data-driven FVC that relies on databases of speakers to estimate empirically the strength of the voice evidence. A key issue for such approaches is the stability of likelihood ratio (LR) output according to the specific speakers used for training and testing systems. This study addresses this issue using simulated scores with different speaker distributions for training and test data. Experiments were replicated 100 times by varying the sampling of (1) both training and test speakers, (2) training speakers only, and (3) test speakers only. The results show that using different speakers for training and testing data affects system stability to different extents, with the C_{llr} varying from 0.51 to 0.61 for the most stable system and 0.03 to 1.46 for the least stable.

Keywords: forensic voice comparison, likelihood-ratio, sampling uncertainty, Bayesian method.

1. INTRODUCTION

Forensic voice comparison (FVC) is a sub-discipline of forensic speech science, which is the application of linguistics, phonetics and acoustics to legal cases [9]. A typical scenario for a FVC case is to compare two recordings, one of an unknown offender (disputed sample), and the other of a known suspect (known sample) typically recorded during the police interview (e.g. in the UK, China) [6] or through wiretaps (e.g. in Germany, China) [12]. The likelihood ratio (LR) framework has been extensively employed and studied in recent years [9,15,21,23]. The LR approach involves evaluating the similarity of the speech patterns in the disputed and known samples and assessing their typicality against a relevant population [8, 10]. The result, which can be expressed using a numerical or verbal LR, is a measure of the strength of the evidence under the competing propositions of the prosecution and defence (for more see [10, 14]). Calculating the LR normally involves two stages: (1) feature-to-score, and (2) score-to-LR.

System performance is widely evaluated by using log LR cost (C_{llr}) [4]. The lower the C_{llr} the more accurate the system is. The term *system* here refers to "a set of procedures and databases that are used to compare two samples, one of known sample and one of disputed sample, and produce a LR" [16]. It is important to evaluate the system performance in order to show how good the system is (i.e. to separate same speaker and different speaker samples). The system evaluation is often carried out by taking a group of speakers (e.g. 60 speakers) and dividing them equally into training, test and background set. The system is then trained, tested and evaluated by using three sets of data. Previous studies have shown that the system performance varies when using matched or mismatched speakers for the relevant background population [e.g. (mis)matched for accent, 8]. Different variables also yield different system accuracy [9,10,15,21,23]. Most of these previous studies have only carried out the experiment once (i.e. with one arrangement of speakers in each dataset). However, relatively little is known about how stable system performance is if a different arrangement of speakers or a different group of 60 speakers is used.

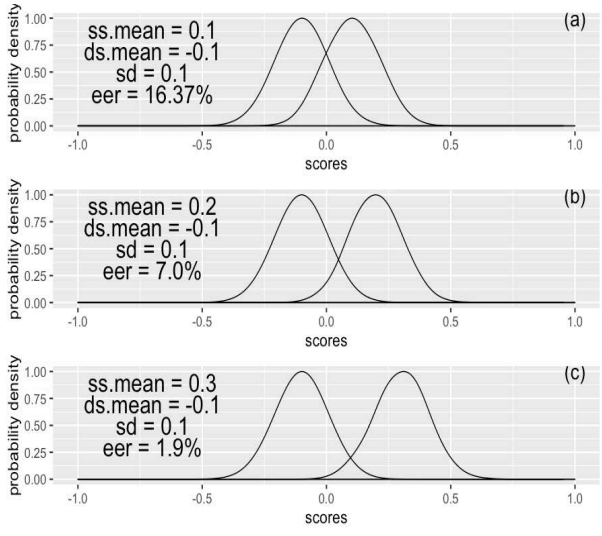
We conducted a study using spontaneous Cantonese speech from 64 speakers to explore the effect of running an experiment multiple times, i.e. by sampling different groups of training, test and background speakers from a relevant population [22]. The results showed that the system performance (C_{llr}s) varied from 0.29 to 1.15 when using different configurations of training, test and background speakers. However, because this study used spontaneous speech, the variability in system stability might have been caused by factors such as number of speakers and tokens used, channel mismatch, and recording qualities of different speakers. Therefore, the current study uses simulated same speaker (SS) and different speaker (DS) scores to address two questions in a controlled manner. First, how is the system stability is affected by sampling, e.g. does the system have a more stable performance if a different set of training, or test, or training and test speakers are used? Second, do some variables provide more or less stable LR output according to the specific sample of speakers used?

2. METHOD

2.1 Data simulation

The data was simulated under a normality assumption. Three sets of simulated scores for 1000 speakers were computed using *rnorm* function in R [1,20], resulting in 1,000 SS and 99,000 DS scores. Note that scores here are numbers simulated from normal distributions and are not computed from actual segmental (e.g. vowels) or suprasegmental (e.g. F0) speech features. Panel (a) in Figure 1 shows the distributions of the simulated SS and DS scores, where the mean and standard deviation of SS scores (right) are 0.1, while the mean and standard deviation of DS scores (left) are -0.1 and 0.1. In panels (b) and (c), the standard deviation of SS and DS scores and the mean of DS scores were kept the same, but the mean of SS scores was increased to 0.2 and 0.3. The three different datasets each had different equal error rates (EER), in order to mimic variables with different speaker-discriminatory power. The data in panel (c) (EER = 1.9%) has the best speaker-discriminatory power; cf. panel (b) (EER = 7%) and panel (a) (EER = 16.37%). This allows us to assess the effect of inherent speaker-discriminatory power on system stability. The three sets of scores were used as the pseudo-datasets for LR computation.

Figure 1: Simulated SS and DS scores with different speaker-discriminatory power.



2.2 LR computation and system evaluation

Since the current study uses simulated scores, only the score-to-LR stage of LR computation is assessed here. Previous studies show that stable LR output can be achieved with 20 or more speakers in each of the training and test data [7]. Therefore, 20 training and test speakers were selected randomly from pseudo-datasets (a), (b) and (c) respectively, which led to 20 SS and 380 DS training and test scores. The training scores were used to generate logistic regression calibration coefficients [3] that were then applied to test scores to produce a set of 20 SS and 380 DS calibrated log LRs. The calibrated C_{llr} was calculated to capture the system performance. The same procedure was repeated 100 times by using *LR calculation and testing in FVC* package [13] in R [2,19]. The overall and interquartile range (IQR) of C_{llr}s are used for system stability evaluation.

3. EXPERIMENT

Three sets of experiments were carried out with pre-defined sampling rules. The speakers were sampled from pseudo-datasets (a), (b) and (c) respectively in each experiment.

3.1 Expt. 1: Sampling training & test speakers.

Different sets of scores was randomly sampled for both training and test data in each replication to explore the effect of speaker-sampling on system stability, and whether some variables produce more or less stable systems according to different samples of speakers used.

3.2 Expt. 2: Only sampling training speakers.

Different sets of scores were randomly sampled for the training speakers while keeping the test scores fixed in each replication. This aims to explore the sensitivity of training data to different speakers with regard to the speaker-discriminatory power of the variable, i.e. to explore whether it matters who we select for the training data if the variable has a higher speaker-discriminatory power, i.e. lower EER.

3.3 Expt. 3: Only sampling test speakers.

Different sets of scores was randomly sampled for test data while the training scores were fixed in each replication. This explores the sensitivity of test data to different speakers and the feasibility of using the same LR-based FVC system for multiple cases.

4. RESULTS

4.1. Experiment 1

Figure 2 shows the variation in C_{llr}s by sampling different sets of training and test scores in each replication. (a), (b) and (c) on the x-axis indicate that the scores were sampled from pseudo-datasets (a), (b) and (c), while the y-axis indicates the C_{llr} values. The

overall C_{llr} ranges from 0.41 to 0.93, 0.08 to 1.06 and 0.01 to 0.94 for sets (a), (b) and (c) respectively. Figure 2 shows firstly that the system stability varies if different sets of SS and DS scores used in each replication. Secondly, sets (a), (b) and (c) yielded different system stabilities. The overall C_{llr} range and IQR of set (b) are larger than those of sets (a) and (c) (Table 1), and the overall C_{llr} range of set (c) is larger than that of set (a). Moreover, set (c) yielded a much lower IQR (0.07, Table 1), and it produced the most outliers. The results show that speaker-sampling has a marked effect on system stability regardless of the speech data being used. Experiments 2 and 3 explore the effects in experiment 1 in more details, to identify whether the training or test data is more important.

Figure 2: Variation of C_{llr}s by sampling training and test speakers from pseudo-datasets (a), (b) and (c).

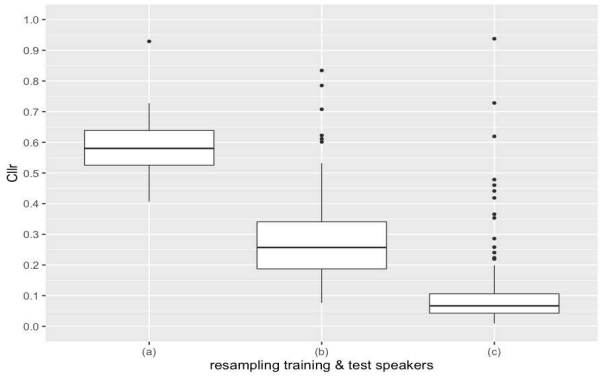


Table 1: minimum, 1st quartile, median, 3rd quartile, maximum and IQR of sets (a), (b) and (c) in experiment 1.

C_{llr}	(a)	(b)	(c)
Min.	0.41	0.08	0.01
1st Qu.	0.52	0.19	0.04
Median	0.58	0.26	0.07
3rd Qu.	0.64	0.34	0.11
Max.	0.93	1.06	0.94
IQR	0.12	0.3	0.07

4.2. Experiment 2

Figure 3 shows the variation in C_{llr}s by sampling different sets of training scores in each replication. One predictable pattern emerges, namely that the further apart the SS and DS scores from each other, the lower the C_{llr} mean and median. However, IQRs of C_{llr} (Table 2) from all three sets are similar, which indicates that variables with higher speaker-discriminatory power do not necessary yield a higher system stability. More outliers are produced when the distributions of SS and DS scores are further apart from each other (c), which makes the overall C_{llr} range of set (c) much higher than (a) and (b).

Figure 3: Variation of C_{llr}s by sampling training speakers from pseudo-datasets (a), (b) and (c).

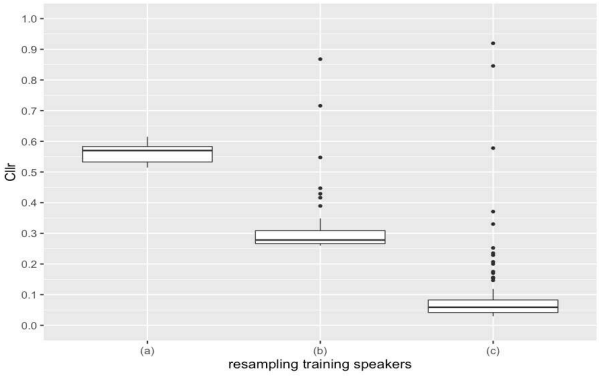


Table 2: minimum, 1st quartile, median, 3rd quartile, maximum and IQR of sets (a), (b) and (c) in experiment 2.

C_{llr}	(a)	(b)	(c)
Min.	0.51	0.26	0.03
1st Qu.	0.53	0.27	0.04
Median	0.57	0.28	0.06
3rd Qu.	0.58	0.31	0.09
Max.	0.61	0.87	1.41
IQR	0.05	0.04	0.05

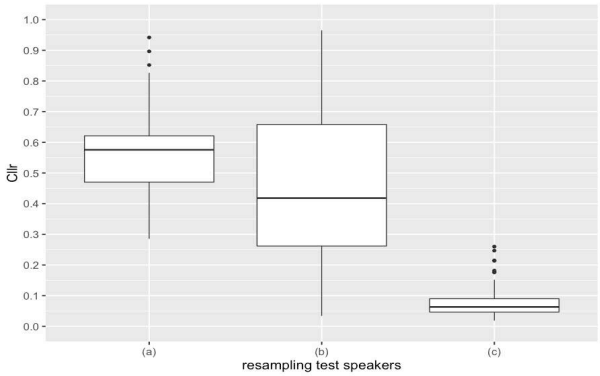
4.3. Experiment 3

Figure 4 shows the variation in C_{llr}s by sampling different sets of scores into test speakers in each replication. The overall range and IQR of C_{llr}s in Experiment 3 yielded a different pattern from Experiment 2. The overall C_{llr}s of set (b) ranges from 0.03 to 1.46 (Table 3) and the IQR is 0.46, which are higher than those of sets (a) and (c). Scores sampled from pseudo-dataset (c) yielded the lowest overall C_{llr} range (0.24) and IQR (0.04), which suggests that it is feasible to use the same LR-based FVC system for multiple FVC caseworks. However, a comparison between sets (a) and (b) shows a different pattern, and it suggests that variable with a higher speaker-discriminatory power does not always yield a higher system stability if different test speakers are used.

Table 3: minimum, 1st quartile, median, 3rd quartile, maximum and IQR of sets (a), (b) and (c) in experiment 3.

C_{llr}	(a)	(b)	(c)
Min.	0.29	0.03	0.02
1st Qu.	0.48	0.28	0.05
Median	0.58	0.44	0.06
3rd Qu.	0.62	0.74	0.09
Max.	1.06	1.46	0.26
IQR	0.14	0.46	0.04

Figure 4: Variation of C_{llr}s by sampling test speakers.



5. DISCUSSION

The results from the three experiments showed that speaker sampling has different effects on system stability.

Experiment 1 shows that sampling both training and test speakers causes the system stability to vary to different extents, and system accuracy is not necessarily positively correlated with system stability. Moreover, the system stability in Experiment 1 might also be related to the calibration method [3] used in the current study. Other calibration methods proposed in [18] might offer a solution to improve system stability.

Experiment 2 shows that sampling training speaker has a limited effect on the system stability regardless of the speaker-discriminatory power of the variables. The IQRs of C_{llr}s of sets (a), (b) and (c) are similar to each other. Moreover, the further away the distributions of SS and DS scores are, the more outliers the system produces, which indicates that variables with higher speaker-discriminatory power does not necessary yield a higher overall system stability.

Experiment 3 shows that sampling test speakers has different effects on system stability for variables with different speaker-discriminatory power. Set (b) in Figure 4 yielded a lower system stability than set (a), which indicates a low feasibility for using the same LR-based FVC for multiple real cases even when the variables have a high speaker-discriminatory power. However, there might be certain thresholds for high or low speaker-discriminatory powers between variables, because set (c) in Figure 4 yielded a much lower variability compared with sets (a) and (b). It is possible that the system starts to yield stable performance when a certain accuracy level is achieved.

Comparison between experiments 1, 2 and 3 shows that sampling training speakers has the least effect on system stability, while sampling test speakers has the most. Comparatively, it matters least which speakers we select for training data. Interestingly, sets (b) yielded lower medians than sets (a) in all three experiments, while the IQR range of sets (b) are no lower than those for sets (a) and (c). This pattern indicates that scores sampled from set (b) is likely to give the worst system stability. It is also apparent that scores sampled from pseudo-dataset (c) consistently yielded the lowest median and IQR across the three experiments. However, sets (c) also gave the most outliers across the three experiments, which suggests the inherent variability of the pseudo-datasets used that leads to the variability in system stability. A potential method to deal with the underlying variability from the input data is to incorporate these types of uncertainty into the LR computation [22] by using fully Bayesian method and Bayesian calibration [5].

6. CONCLUSION

The current study used simulated data to explore the effect of speaker sampling on the system stability. The results reinforced the underlying uncertainty in data-driven speech comparison studies and have few implications for both LR-based FVC and phonetic studies in general. Firstly, it is necessary to capture both system accuracy and stability rather than reporting one single C_{llr} value in LR-based FVC. Secondly, the variability in source data causes the system performance to vary to different extents regardless of the speaker-discriminatory power of the variables being used, namely variables with higher speaker-discriminatory power do not necessarily yield higher system stability. Thirdly, it is essential to replicate experiment multiple times. Otherwise, the results would be misleading in the subsequent court ruling, facing risks of convicting an innocent man or setting guilty free.

7. REFERENCES

[1] Ahrens, J. H., Dieter, U. (1973). Extensions of Forsythe's method for random sampling from the normal distribution. *Mathematics of Computation*, **27**, 927–937.

[2] Aitken, C. G., Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109-122.

[3] Brümmer, N. et al. (2007) Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing, 15*, pp. 2072-2084.

[4] Brümmer, N., Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2-3), 230-275.

[5] Brümmer, N., Swart, A. (2014). Bayesian calibration for foresicn evidence reporting. *Proc. Interspeech 2014.* Singapore pp. 388-392.

[6] Home Office. (2003). Criminal Justice Act (Chapter 44). Her Majesty's Stationery Office.

[7] Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15-29.

[8] Hughes, V., Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, *66*, 218-230.

[9] Hughes, V., Foulkes, P., Wood, S. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, volume 99-132.

[10] Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, *2*(4), 671-711.

[11] Kinoshita, Y., Ishihara, S., Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language & the Law*, *16*(1).

[12] Liu, X. M. (2006). Criminal investigation theory and reform 刑事侦查程序理论与改革研究. China Legal Publishing House.

[13] Lo, J. (2018). fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison [unpublished R package] version 0.1.0.

[14] Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298-308.

[15] Morrison, G. S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*(4), 2387-2397.

[16] Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173-197.

[17] Morrison, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science & Justice*, *56*(5), 371-373.

[18] Morrison, G. S., Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, 58(3), 200-218.

[19] Morrison, G.S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. [Software].

[20] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for statistical Computing, Vienna, Austria. http://www.R-project.org/

[21] Rose, P., Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Odyssey 2016*, 326-333.

[22] Wang, B. X., Hughes, V., Foulkes, P. (2018) A preliminary investigation of the effect of speaker randomisation in likelihood-ratio based forensic voice comparison. *IAFPA 2018*. University of Huddersfield, 125-125.

[23] Zhang, C., Morrison, G. S., Thiruvaran, T. (2011). Forensic voice comparison using Chinese/iau/. *Proc. 17th ICPhS* Hong Kong, 21.